A Constrained Maximum Likelihood Estimator of Speech and Noise Spectra with Application to Multi-Microphone Noise Reduction

Zahedi, Adel; Pedersen, Michael; Østergaard, Jan; Bramsløw, Lars; Christiansen, Thomas; Jensen, Jesper

Published in:
ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI (link to publication from Publisher):
10.1109/ICASSP40776.2020.9053077

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

A CONSTRAINED MAXIMUM LIKELIHOOD ESTIMATOR OF SPEECH AND NOISE SPECTRA WITH APPLICATION TO MULTI-MICROPHONE NOISE REDUCTION

Adel Zahedi[†*], *Michael Syskind Pedersen*[†], *Jan Østergaard*[*], *Lars Bramsløw*[†],
Thomas Ulrich Christiansen[†], *Jesper Jensen*[†*]
[†] Oticon A/S, Smørum, Denmark
[*] Aalborg University, Aalborg, Denmark

ABSTRACT

One of the challenges with the implementation of multi-microphone noise reduction systems in practical applications lies in the need for the knowledge of the speech and noise covariance matrices. Recently, a method based on Maximum Likelihood (ML) estimation addressed this problem. Despite its relative success in practical setups, this method may suggest negative spectral components for the clean speech due to noise influences. In this paper, we suggest a new estimation technique that tackles this issue by enforcing a power constraint on the estimation problem. We compare the proposed method with the ML method both in synthetic and real-life scenarios using objective measures. The results suggest that the proposed method can improve speech quality without a loss of intelligibility.

Index Terms— Multichannel Wiener filter, maximum likelihood estimation, hearing-assistive devices, water filling

1. INTRODUCTION

Poor performance in noise is one of the most common points of dissatisfaction for the users of hearing-assistive devices (HADs) [1, 2]. Noise reduction is for this reason an integral part of most modern HADs. One of the most well-known noise reduction techniques is the multi-channel Wiener filter (MWF) [3, 4]. Despite offering simple closed-form solutions, implementation of MWFs in practical setups such as in HADs tangles with practicalities, among which estimation of the generally time-varying inter-microphone statistics of speech and noise is particularly challenging.

The MWF can be decomposed as a cascade of an MVDR beamformer and a single-channel postfilter [5]. Several methods have been proposed for estimating the signal statistics necessary to implement the MWF beamformer in general [6–10] and the speech and noise power spectral densities (PSD) for implementing the postfilter in particular [11–14]. In [13], a Maximum Likelihood (ML) scheme was proposed for estimating the speech and noise PSDs. This method has been successfully used for scientific [15] as well as industrial [14, 16] applications. However, typically there are some frequency bins where the ML estimation scheme suggests negative values for the speech spectrum. Rounding these components up to zero, which is often done in practical speech enhancement systems [17], leads to an overall tendency to overestimate the speech power (cf. Section 2 for more details). In this paper, we propose an estimation technique that alleviates this issue. Although the proposed method can be applied for speech an noise PSD estimation in a broader context, in this paper, we focus on noise reduction using the MWF.

The remaining of this paper is organized as follows: In Section 2, we review the ML estimation method of [13]. In Section 3, we propose a new formulation and solve the resulting problem. In Section 4, we compare the proposed method with the one in [13] using simulations in a hearing aid setup. Section 5 concludes the paper.

This work was partially funded by Innovation Fund Denmark.

2. ML ESTIMATION OF SPEECH AND NOISE SPECTRA

In the short-time Fourier transform domain, we use the following model for the noisy speech acquired by M microphones:

$$\mathbf{X}(k,l) = S(k,l)\mathbf{d}(k,l) + \mathbf{V}(k,l), \tag{1}$$

where the M-dimensional vectors $\mathbf{X}(k,l)$ and $\mathbf{V}(k,l)$ respectively represent noisy speech and noise signals at the M microphones at frequency bin k and time frame l. The clean speech signal at the reference microphone is denoted by $S(k,l)$, and the M-dimensional vector $\mathbf{d}(k,l)$ is the relative transfer function for the M microphones [18]; i.e. the transfer function from the target speech source to the M microphones normalized by the one for the reference microphone. Assuming that the noise and speech signals are uncorrelated and using (1), the covariance matrix of the noisy speech is given by:

$$\mathbf{C}_x(k,l) = \lambda_s(k,l)\mathbf{d}(k,l)\mathbf{d}^H(k,l) + \lambda_v(k,l)\mathbf{\Gamma}(k,l), \tag{2}$$

where $\lambda_s(k,l) = |S(k,l)|^2$ and $\lambda_v(k,l)$ are, respectively, the clean speech and noise spectra at the reference microphone, and $\mathbf{\Gamma}(k,l)$ is the noise covariance matrix normalized by the noise variance at the reference microphone. One can say that $\mathbf{\Gamma}(k,l)$ represents the *structure* of the noise covariance matrix. Using a voice activity detector, the noise covariance matrix can be estimated directly during the speech absence intervals. Assuming that the structure of the covariance matrix remains unchanged during speech activity intervals, (2) can be written as:

$$\mathbf{C}_x(k,l) = \lambda_s(k,l)\mathbf{d}(k,l)\mathbf{d}^H(k,l) + \lambda_v(k,l)\mathbf{\Gamma}(k,l_0), \tag{3}$$

where l_0 indexes the most recent frame with no speech activity. Given that the relative transfer functions $\mathbf{d}(k,l)$ are known, the only unknown parameters left in (3) are $\lambda_s(k,l)$ and $\lambda_v(k,l)$. Assume that $\mathbf{X}(k,l)$ follows a zero-mean complex circularly symmetric Gaussian distribution with the covariance matrix given in (3); i.e.

$$f_X\left(\mathbf{X}(k,l); \lambda_s(k,l), \lambda_v(k,l)\right) = \mathcal{CN}(\mathbf{0}, \mathbf{C}_x(k,l)). \tag{4}$$

Also suppose that D independent observations of the noisy speech are available; e.g. D consecutive frames $\underline{\mathbf{X}}_D(k,l) = [\mathbf{X}(k,l-D+1)\dots\mathbf{X}(k,l)]$ assuming independence across the frames. The joint probability density function (pdf) of $\underline{\mathbf{X}}_D(k,l)$ is simply given by the product of the density functions of the individual frames, and the ML estimation of $\lambda_s(k,l)$ and $\lambda_v(k,l)$ can be obtained by maximizing the resultant joint pdf; i.e.

$$\max_{\lambda_s(k,l),\lambda_v(k,l)} \ln f_{\underline{X}_D}(\underline{\mathbf{X}}_D(k,l); \lambda_s(k,l), \lambda_v(k,l)), \tag{5}$$

which can be solved in closed-form, yielding the following [13, 19]:

$$\lambda_v^{ML}(k,l) = \frac{1}{M-1}\text{tr}\left(\frac{1}{D}\underline{\mathbf{X}}_D^H(k,l)\mathbf{B}(k,l)\times \right.$$
$$\left.\left(\mathbf{B}^H(k,l)\mathbf{\Gamma}(k,l_0)\mathbf{B}(k,l)\right)^{-1}\mathbf{B}^H(k,l)\underline{\mathbf{X}}_D(k,l)\right), \quad (6)$$

$$\lambda_s^{ML}(k,l) = \mathbf{w}^H(k,l)\left(\hat{\mathbf{C}}_x(k,l) - \lambda_v^{ML}(k,l)\mathbf{\Gamma}(k,l_0)\right)\mathbf{w}(k,l), \quad (7)$$

where the $M \times M-1$ blocking matrix $\mathbf{B}(k,l)$ can be calculated as the first $M-1$ columns of $\mathbf{I}_M - \mathbf{d}(k,l)\mathbf{d}^H(k,l)/\mathbf{d}^H(k,l)\mathbf{d}(k,l)$, and $\hat{\mathbf{C}}_x(k,l)$ and $\mathbf{w}(k,l)$ (the MVDR beamformer weight vector) are defined as:

$$\hat{\mathbf{C}}_x(k,l) \triangleq \frac{1}{D}\sum_{j=l-D+1}^{l}\mathbf{X}(k,j)\mathbf{X}^H(k,j), \quad (8)$$

$$\mathbf{w}(k,l) \triangleq \frac{\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}{\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}. \quad (9)$$

The estimator given by (6) and (7) is the minimum-variance unbiased estimator, thus achieving the Cramér-Rao lower bound [13]. However, when the noise level is large compared to the speech level at a certain frequency bin k, $\lambda_s^{ML}(k,l)$ in (7) may become negative. This can happen even at high global SNRs at frequency bins where the speech power is low. The typical treatment in such cases is to round up the negative values to zero (equivalent to adding a nonnegativity constraint to (5)). However, one can argue that as the negative values of $\lambda_s^{ML}(k,l)$ are due to the noise influence, there is no reason to believe that the positive ones are not, especially taking into account that the estimator is unbiased. Getting rid of the negative values by trimming them to zero at some frequency bins, leaves us with spurious positive estimates at some other frequency bins, which give rise to a net effect of overestimating the speech power. Consequently, when used in an MWF context, the noise in the resulting enhanced speech signal would be under-suppressed. In the next Section, we propose a method that addresses this issue.

3. PROPOSED METHOD

3.1. Problem Formulation

Suppose that $\mathcal{K} = \{1, \ldots, K\}$ is the set of all frequency bins. Problem (5) is defined over individual frequency bins, and one needs to solve it separately for each and every $k \in \mathcal{K}$. Equivalently, one can write the joint pdf for all frequency bins as the product of the individual pdfs in (4), and obtain the same solution as in (6)–(7) by solving the following problem:

$$\max_{\substack{\lambda_s(1,l),\ldots,\lambda_s(K,l)\\ \lambda_v(1,l),\ldots,\lambda_v(K,l)}} \ln\prod_{k=1}^{K} f_{\underline{X}_D}(\underline{\mathbf{X}}_D(k,l); \lambda_s(k,l), \lambda_v(k,l)). \quad (10)$$

As argued in Section 2, when the noise influence is significant, $\lambda_s^{ML}(k,l)$ resulting from (10) may take negative or positive spurious values depending on the frequency bin. Let us denote the ML estimate of the speech power in frame l by $P_s^{ML}(l)$; i.e.

$$P_s^{ML}(l) \triangleq \sum_{k=1}^{K}\lambda_s^{ML}(k,l). \quad (11)$$

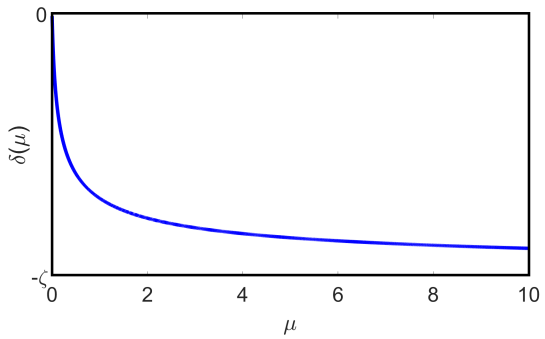

Fig. 1. An illustration of the water-filling solution.

Note that $P_s^{ML}(l)$ averages the noise influence over the individual spectral components, and is therefore likely to be less noisy than the individual estimates $\lambda_s^{ML}(k,l)$. Based on this rationale, we introduce a power constraint to (10) to formulate a new estimation problem as follows:

$$\max_{\substack{\lambda_s(1,l),\ldots,\lambda_s(K,l)\\ \lambda_v(1,l),\ldots,\lambda_v(K,l)}} \ln\prod_{k=1}^{K} f_{\underline{X}_D}(\underline{\mathbf{X}}_D(k,l); \lambda_s(k,l), \lambda_v(k,l))$$
$$\text{s.t.} \quad \sum_{k=1}^{K}\lambda_s(k,l) = P_s^{ML}(l) \quad \text{and} \quad \lambda_s(k,l) \geq 0 \text{ for all } k \in \mathcal{K}. \quad (12)$$

3.2. Problem Solution

We prove in Appendix that the solution to (12) for $\lambda_v(k,l)$ is the same as $\lambda_v^{ML}(k,l)$ as expected (since the constraint in (12) does not depend on $\lambda_v(k,l)$), and for $\lambda_s(k,l)$ it is given in the following water-filling form:

$$\lambda_s^*(k,l) = \left(\lambda_s^{ML}(k,l) - \zeta(k,l) + \frac{\sqrt{2\zeta(k,l)\mu(l)+1}-1}{\mu(l)}\right)^+, \quad (13)$$

where $(\cdot)^+ \triangleq \max(\cdot, 0)$, $\zeta(k,l) \triangleq \mathbf{w}^H(k,l)\hat{\mathbf{C}}_x(k,l)\mathbf{w}(k,l)$, and the water level $\mu(l) \geq 0$ is adjusted such that the following holds:

$$\sum_{k=1}^{K}\lambda_s^*(k,l) = P_s^{ML}(l). \quad (14)$$

The water level $\mu(l) \geq 0$ can be calculated using any of the available efficient algorithms (cf. [20]) or simply using bisection. The graph of the term $\delta(\mu) \triangleq -\zeta + (\sqrt{2\zeta\mu+1}-1)/\mu$ is shown in Fig. 1. When $\lambda_s^{ML}(k,l) \geq 0$ for all $k \in \mathcal{K}$, the water level is $\mu(l) = 0$, yielding $\lambda_s^*(k,l) = \lambda_s^{ML}(k,l)$. When $\lambda_s^{ML}(k,l) < 0$ for at least one k, $\delta(\mu)$ is always negative, implying that $\lambda_s^*(k,l) < \lambda_s^{ML}(k,l)$. This however is only the case in frequency bins where $\lambda_s^{ML}(k,l) > 0$. In other bins, the $(\cdot)^+$ operator in (13) sets $\lambda_s^*(k,l)$ equal to 0. In summary, to calculate $\lambda_s^*(k,l)$ from $\lambda_s^{ML}(k,l)$, all negative components are trimmed to 0, and each positive one is reduced by an amount that depends on its corresponding $\zeta(k,l)$.

3.3. Subband Implementation

The MWF is optimal in sense of mean-squared error (MSE). The proposed method may lead to an implementation that is closer to the

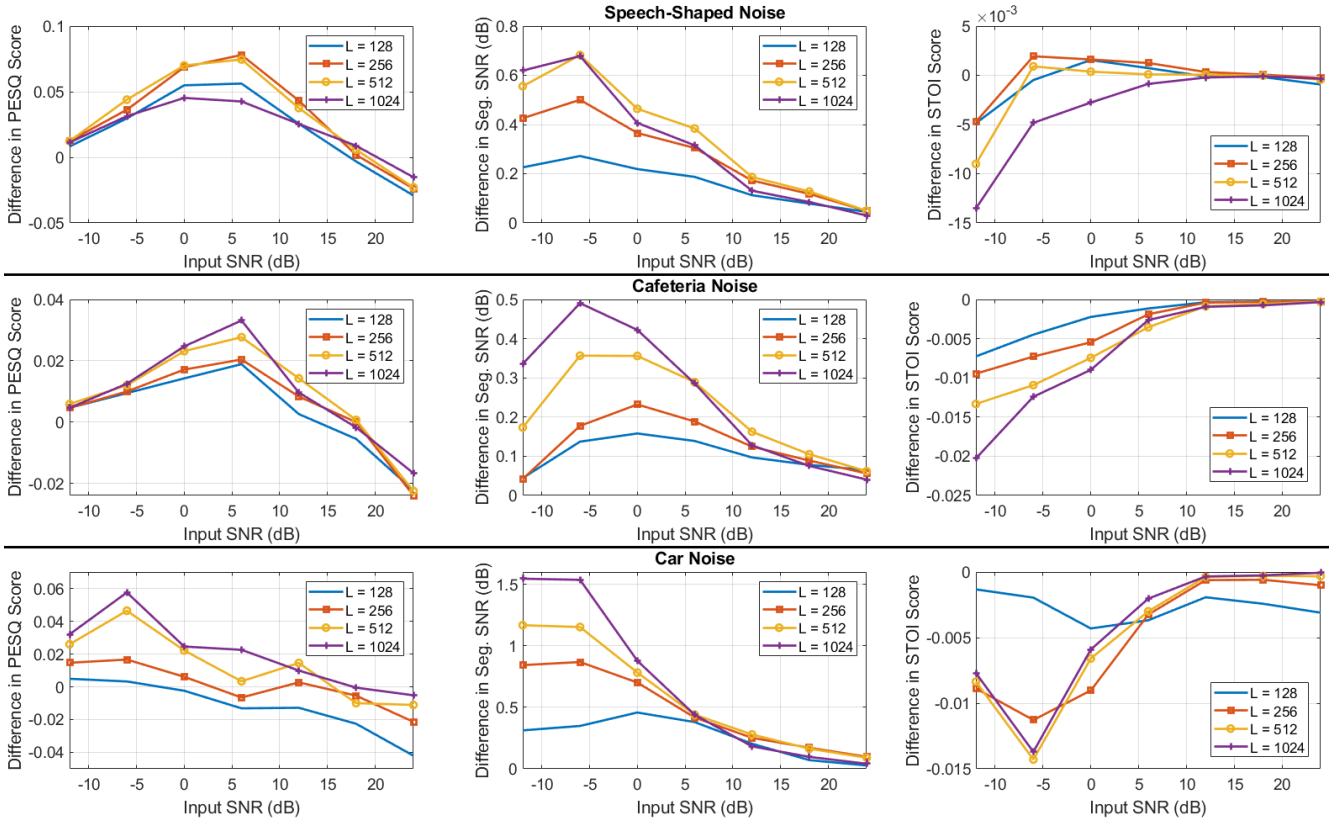

Fig. 2. Improvement gained by the proposed method over the ML estimation with three objective performance measures, four different frame sizes and three different noise types.

Table 1. Scores for the ML method with $L = 256$ in cafeteria noise.

Input SNR	-12	-6	0	6	12	18	24
PESQ	1.06	1.18	1.42	2.00	2.58	3.25	3.60
Seg. SNR	-6.62	0.40	6.10	12.05	17.36	22.63	28.90
STOI	0.57	0.70	0.81	0.90	0.96	0.98	0.99

ideal MWF, yielding a lower MSE. This, however, may not necessarily translate into perceptual improvements. To optimize the performance in a more perceptual-oriented manner, we implement (12) in subbands. Suppose that $k_i^{\min}$ and $k_i^{\max}$ index the lowest and highest frequencies in subband i, respectively. The ML estimation of the speech power in subband i is given by:

$$P_{s,i}^{ML}(l) \triangleq \sum_{k=k_i^{\min}}^{k_i^{\max}} \lambda_s^{ML}(k,l). \tag{15}$$

Obviously, $\lambda_s^*(k,l)$ will still be given by (13), but the subband-dependent water level $\mu_i(l)$ should be adjusted such that the following holds:

$$\sum_{k=k_i^{\min}}^{k_i^{\max}} \lambda_s^*(k,l) = P_{s,i}^{ML}(l). \tag{16}$$

4. SIMULATION RESULTS

We compare the performance of an MWF that operates using the proposed method against one that uses the ML estimation scheme (6)–(7) both with synthetic and realistic noises. For synthetic noise, we create stationary Gaussian speech-shaped noise (SSN) by modelling the long-term spectrum of the target speech using an LPC model of order 15. The SSN is then convolved with a set of head-related transfer functions (HRTF) measured in a setup, where a dummy head is located at the center of a circular array of 48 loudspeakers with an angular resolution of 7.5 degrees, and a hearing aid with $M = 2$ microphones is placed on the left ear. As for the realistic noise, we use sound recordings in two realistic scenes using a spherical array of 32 microphones: one in Oticon's cafeteria during the lunch hours and the other inside the cabin of a car driving steadily on a highway. The recorded sound field is then recreated in a room using three circular uniform arrays of loudspeakers at elevations of -45, 0 and 45 degrees with respect to a dummy head placed at the center, and consisting of 6, 16 and 6 loudspeakers, respectively. The HRTFs from each loudspeaker to each of the $M = 2$ microphones of a hearing aid placed on the left ear of the dummy head are measured and used in the simulations. The target speech is assumed to be frontal, and is randomly selected from the TIMIT test set [21] prepended with a silence interval of duration 0.2 seconds during which $\mathbf{\Gamma}(k,l_0)$ is estimated. In all simulations, we set $D = 5$ and $G_{\min} = -7$ dB, where $G_{\min}$ is the minimum gain for the single-channel postfilter.

The proposed method is implemented in octave bands with the upper frequency limits of 0.25, 0.5, 1, 2, 4 and 10 kHz. The SNR at the reference microphone is varied from -12 to +24 dB with a 6 dB stepsize. At each SNR, 25 speakers are randomly chosen from the dataset and one sentence is randomly picked from each speaker to perform one independent trial.

We use three objective measures to compare the two methods: Perceptual Evaluation of Sound Quality (PESQ) [22], Segmental SNR, and the Short-Time Objective Intelligibility (STOI) [23]. The improvements gained by the proposed method over the ML estimation method are shown in Fig. 2 for various frame sizes L. The corresponding absolute scores for the ML method for $L = 256$ and cafeteria noise are summarized in Table 1 for reference. As the plots suggest, in a range of SNRs that are of most interest for noise reduction systems, the performance is improved by the proposed method in sense of PESQ and Segmental SNR. In sense of estimated intelligibility, the difference in the STOI scores is mostly insignificant. Thus it is reasonable to say the improvements in speech quality are obtained without compromising the intelligibility.

5. CONCLUSIONS

We addressed the problem of negative components in the estimated speech spectrum, when using ML estimation in a Multichannel Wiener Filter context. We proposed a new formulation, which modifies the ML approach by making sure that the estimated speech power is not biased due to negative components. We solved the new problem and obtained a water filling type of solution. We compared the proposed method with the ML method in both synthetic and realistic environments using objective measures. The proposed method improves the performance in sense of PESQ and segmental SNR in a range of SNRs that are of practical significance, while in terms of estimated intelligibility (STOI) the performance remains unchanged or changes by insignificant amounts.

6. APPENDIX

The unconstrained version of (12) (the ML estimation) can be solved for $\lambda_v(k,l)$ without entanglement with $\lambda_s(k,l)$ [19]. The solution is given in (6). Since the constraint in (12) does not depend on $\lambda_v(k,l)$, the same approach will solve (12) for $\lambda_v(k,l)$, yielding the same solution. In the sequel, we substitute this solution in (12) and solve the resultant problem for $\lambda_s(k,l)$. We make use of the following identities [24]: (a) $\partial \ln |\mathbf{A}| = \text{tr}\left(\mathbf{A}^{-1}\partial\mathbf{A}\right)$, where tr$(\cdot)$ is the trace operator, (b) $\partial\text{tr}\left(\mathbf{A}\right) = \text{tr}\left(\partial\mathbf{A}\right)$, (c) $\text{tr}\left(\mathbf{ABC}\right) = \text{tr}\left(\mathbf{CAB}\right)$, and (d) $\frac{\partial \mathbf{A}^{-1}}{\partial x} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$, where x is a real scalar.

Disregarding the nonnegativity constraint in (12) for now, we write it in Lagrangian form (with the Lagrange multiplier γ) and set its derivative w.r.t. $\lambda_s(k,l)$ equal to zero to obtain:

$$\frac{\partial}{\partial\lambda_s(k,l)}\left(-D\ln|\mathbf{C}_x(k,l)| - \sum_{j=l-D+1}^{l}\mathbf{X}^H(k,l)\mathbf{C}_x^{-1}(k,l)\mathbf{X}(k,l)\right) - \gamma = 0. \tag{17}$$

We apply the Matrix Inversion Lemma [25] to (3) to obtain:

$$\mathbf{C}_x^{-1}(k,l) = \frac{1}{\lambda_v^{ML}(k,l)}\times \left(\mathbf{\Gamma}^{-1}(k,l_0) - \frac{\lambda_s(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)}\right), \tag{18}$$

where

$$a(k,l) \triangleq \mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l) = \frac{1}{\mathbf{w}^H(k,l)\mathbf{\Gamma}(k,l_0)\mathbf{w}(k,l)}, \tag{19}$$

where the last equality in (19) follows from (9). Using (18), identities (a)–(c) and (3), we have:

$$\begin{aligned}
&\partial\ln|\mathbf{C}_x(k,l)| = \text{tr}\left(\mathbf{C}_x^{-1}(k,l)\partial\mathbf{C}_x(k,l)\right) = \frac{1}{\lambda_v^{ML}(k,l)}\times \\
&\text{tr}\left[\left(\mathbf{\Gamma}^{-1}(k,l_0) - \frac{\lambda_s(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)}\right)\partial\mathbf{C}_x(k,l)\right] \\
&= \frac{1}{\lambda_v^{ML}(k,l)}\partial\left[\text{tr}\left(\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{C}_x(k,l)\right)\right] - \\
&\frac{1}{\lambda_v^{ML}(k,l)}\frac{\lambda_s(k,l)\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\left(\partial\mathbf{C}_x(k,l)\right)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)} \\
&= \frac{1}{\lambda_v^{ML}(k,l)}\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)\partial\lambda_s(k,l) - \\
&\frac{\lambda_s(k,l)}{\lambda_v^{ML}(k,l)}\frac{\left(\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)\right)^2\partial\lambda_s(k,l)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)} \\
&= \frac{a(k,l)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)}\partial\lambda_s(k,l).
\end{aligned} \tag{20}$$

Applying identity (d) and then using (3), we have:

$$\begin{aligned}
&\frac{\partial}{\partial\lambda_s(k,l)}\mathbf{X}^H(k,j)\mathbf{C}_x^{-1}(k,j)\mathbf{X}(k,j) = \mathbf{X}^H(k,j)\frac{\partial\mathbf{C}_x^{-1}(k,j)}{\partial\lambda_s(k,l)}\mathbf{X}(k,j) \\
&= -\mathbf{X}^H(k,j)\mathbf{C}_x^{-1}(k,j)\mathbf{d}(k,l)\mathbf{d}^H(k,l)\mathbf{C}_x^{-1}(k,j)\mathbf{X}(k,j).
\end{aligned} \tag{21}$$

Substituting (18) in (21) and simplifying the result yields:

$$\begin{aligned}
&\frac{\partial}{\partial\lambda_s(k,l)}\mathbf{X}^H(k,j)\mathbf{C}_x^{-1}(k,j)\mathbf{X}(k,j) \\
&= -\frac{\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{X}(k,j)\mathbf{X}^H(k,j)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}{\left(\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)\right)^2}.
\end{aligned} \tag{22}$$

Substituting (22) and (20) in (17) and using (8), we obtain:

$$\begin{aligned}
&-\frac{a(k,l)}{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)} + \\
&\frac{\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\hat{\mathbf{C}}_x(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}{\left(\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)\right)^2} - \frac{\gamma}{D} = 0.
\end{aligned} \tag{23}$$

Rearranging the terms in (23), we rewrite it as follows:

$$\begin{aligned}
&\frac{\mathbf{d}^H(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\hat{\mathbf{C}}_x(k,l)\mathbf{\Gamma}^{-1}(k,l_0)\mathbf{d}(k,l)}{a^2(k,l)} - \frac{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)}{a(k,l)} \\
&- \frac{\gamma}{D}\left(\frac{\lambda_v^{ML}(k,l)+a(k,l)\lambda_s(k,l)}{a(k,l)}\right)^2 = 0.
\end{aligned} \tag{24}$$

Combining (19) and (7), we have $\lambda_v^{ML}(k,l) + a(k,l)\lambda_s^{ML}(k,l) = a(k,l)\zeta(k,l)$. Using this and (7), we rewrite (24) as the following second order polynomial equation:

$$\left(\lambda_s(k,l) - \lambda_s^{ML}(k,l)\right) + \frac{\gamma}{D}\left(\zeta(k,l) + \lambda_s(k,l) - \lambda_s^{ML}(k,l)\right)^2 = 0. \tag{25}$$

Solving (25) for $\lambda_s(k,l)$ and discarding the unfeasible root yields:

$$\lambda_s^*(k,l) = \left(\lambda_s^{ML}(k,l) - \zeta(k,l) + \frac{\sqrt{4\zeta(k,l)\frac{\gamma}{D}+1}-1}{2\frac{\gamma}{D}}\right). \tag{26}$$

Defining $\mu \triangleq 2\frac{\gamma}{D}$ and applying the $(\cdot)^+$ operator to (26) to take account of the nonnegativity constraint in (12) gives (13). This together with the power constraint in (12) yield the water-filling solution.

7. REFERENCES

[1] A. Boothroyd, K. Fitz, J. Kindred, S. Kochkin, H. Levitt, B. Moore, and J. Yanz, "Hearing aids and wireless technology," *Hearing Review*, vol. 14, no. 6, p. 44, 2007.

[2] N. Bisgaard and S. Ruf, "Findings from eurotrak surveys from 2009 to 2015: Hearing loss prevalence, hearing aid adoption, and benefits of hearing aid use," *American Journal of Audiology*, vol. 26, no. 3S, pp. 451–461, 2017.

[3] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[4] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," 2003.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[7] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[8] M. Taseska and E. Habets, "Nonstationary noise psd matrix estimation for multichannel blind speech extraction," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 11, pp. 2223–2236, 2017.

[9] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1136–1150, 2019.

[10] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 544–548.

[11] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 6, pp. 1052–1067, 2018.

[12] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood psd estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, 2016.

[13] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5728–5732.

[14] J. Jensen, M. S. Pedersen *et al.*, "Microphone system and a hearing device comprising a microphone system," Dec. 13 2018, uS Patent App. 16/003,396.

[15] A. I. Koutrouvelis, J. Jensen, M. Guo, R. C. Hendriks, and R. Heusdens, "Binaural speech enhancement with spatial cue preservation utilising simultaneous masking," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 598–602.

[16] A. T. Bertelsen, M. S. Pedersen, J. Jensen, T. Kaulberg, and M. Christophersen, "Hearing device comprising a beamformer filtering unit," Oct. 12 2017, uS Patent App. 15/482,188.

[17] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[18] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[19] H. Ye and D. DeGroat, "Maximum likelihood doa estimation and asymptotic cramér-rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[20] D. Pérez Palomar and J. Rodríguez Fonollosa, "Practical algorithms for a family of waterfilling solutions," 2005.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[24] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*. http://matrixcookbook.com, 2012.

[25] D. S. Bernstein, *Matrix mathematics: theory, facts, and formulas*. Princeton university press, 2009.